Lab 14: Bootstrap Confidence Intervals

Hesketh and Everitt (2000) report on a study by Caplehorn and Bell (1991) that investigated the times (in days) that heroin addicts remained in a clinic for methadone maintenance treatment. The data read in below include the amount of time that the subjects stayed in the facility until treatment was terminated (`time`). For about 37% of the subjects, the study ended while they were still the in clinic (`status`==0). Thus, their survival time has been truncated. For this reason we might not want to estimate the mean survival time, but rather some other measure of typical survival time. Below we explore using the median. We treat the group of 238 patients as representative of the population.

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
heroin <- read_csv("http://www.evanlray.com/data/misc/heroin.txt")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   clinic = col_double(),
##   status = col_double(),
##   time = col_double(),
##   prison = col_double(),
##   dose = col_double()
## )
```

```
head(heroin)
```

```
## # A tibble: 6 x 6
##      id clinic status  time prison  dose
##   <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl>
## 1     1      1      1   428      0    50
## 2   132      2      0   633      0    70
## 3     2      1      1   275      1    55
## 4   133      2      1   661      0    40
## 5     3      1      1   262      0    55
## 6   134      2      1   232      1    70
```

```
nrow(heroin)
```

```
## [1] 238
```

1. Obtain a 95% bootstrap percentile confidence interval for the population median

You could use $B = 10000$ bootstrap samples.

Your code should do something like the following steps:

1. Allocate space to store your results

2. For b = 1, . . . , 10000
    a. Draw a bootstrap sample of size $n = 238$ from the observed data with replacement
    b. Compute the median of the bootstrap sample from step 2a, and save it
3. Find the percentiles of the resulting distribution

```
# number of observations in sample_obs
n <- nrow(heroin)

# how many bootstrap samples to take, and storage space for the results
num_bootstrap_samples <- 10^3
bootstrap_estimates <- data.frame(
  estimate = rep(NA, num_bootstrap_samples)
)

# draw many samples from the observed data and calculate mean of each simulated sample
for(i in seq_len(num_bootstrap_samples)) {
  ## Draw a bootstrap sample of size n with replacement from the observed data
  bootstrap_resampled_obs <- heroin %>%
    sample_n(size = n, replace = TRUE)

  ## Calculate mean of bootstrap sample
  bootstrap_estimates$estimate[i] <- median(bootstrap_resampled_obs$time)
}

quantile(bootstrap_estimates$estimate, probs = c(0.025, 0.975))
```

```
##    2.5%   97.5%
## 319.500 450.025
```

2. Obtain a 95% bootstrap t confidence interval for the population median

We don't have a convenient formula for the standard error of the population median, so you will have to use a nested bootstrap, with the inner loop used to obtain the bootstrap standard error of the median. This will take a long time to run, so while you are developing your code set B to a small number like 10. Once you are confident that your code is working, you can up the number of bootstrap iterations to 1000 or so.

```
# number of observations in sample_obs
n <- nrow(heroin)

# how many bootstrap samples to take, and storage space for the results
num_bootstrap_samples <- 10^3
bootstrap_ts <- data.frame(
  t = rep(NA, num_bootstrap_samples)
)

bootstrap_medians <- data.frame(
  median = rep(NA, num_bootstrap_samples)
)

inner_bootstrap_medians <- data.frame(
  median = rep(NA, num_bootstrap_samples)
)

# draw many samples from the observed data and calculate mean of each simulated sample
for(i in seq_len(num_bootstrap_samples)) {
  ## Draw a bootstrap sample of size n with replacement from the observed data
  bootstrap_resampled_obs <- heroin %>%
    sample_n(size = n, replace = TRUE)

  for(j in seq_len(num_bootstrap_samples)) {
```

```
    inner_resample <- bootstrap_resampled_obs %>%
      sample_n(size = n, replace = TRUE)
    inner_bootstrap_medians$median[j] <- median(inner_resample$time)
  }

  ## Calculate t statistic based on bootstrap sample
  bootstrap_ts$t[i] <- (median(bootstrap_resampled_obs$time) - median(heroin$time)) /
    sd(inner_bootstrap_medians$median)

  ## Calculate median based on bootstrap sample
  bootstrap_medians$median[i] <- median(bootstrap_resampled_obs$time)
}

median(heroin$time) -
  quantile(bootstrap_ts$t, probs = c(0.975, 0.025)) * sd(bootstrap_medians$median)
```

```
##    97.5%     2.5%
## 296.7801 419.1258
```